Exceptions to the Rule

At least two human languages contain grammatical features that put them outside some well-known grammars

By IVARS PETERSON

In the African language Bambara, the phrase *wuluniyninanyinina o wulunyininanyinina* means "whoever searches for dog searchers." Although this expression doesn't slip smoothly into typical cocktail party chatter, the recent recognition of this type of grammatical construction settles a longstanding linguistic dispute. At issue is the nature of the simplest grammer that encompasses all human languages.

"Striking discoveries about human languages are not very common in science," says linguist Geoffrey K. Pullum of the University of California at Santa Cruz. "But this is a nice instance of a genuinely new discovery about natural languages."

The finding may even have implications for programming computers to understand human conversation (SN: 7/27/85, p. 53). "It's now [recognized to be] a more demanding task," says Pullum.

For decades, linguists have delved into words and sentences, pondering the subtle interplay of position, meaning and pronunciation that adds up to a meaningful, grammatical sentence. Part of this study has been an attempt to pick out the hidden framework that underlies human languages and allows them to work as effective message carriers.

In any human language, there are certain things that native speakers, even without the prompting of a schoolteacher, naturally and readily identify as grammatical. These expressions are part of the language, while other, "ungrammatical" sentences clearly don't belong.

This ability to separate grammatical from ungrammatical sentences provides some of the data that go into the study of language structure. But it isn't enough just to observe what people say and write. Linguists also try to make some sense out the observations and to explain why languages seem to be structured in particular ways.

"You wind up hypothesizing a lot of theoretical machinery that's not directly reflected in any way in the observed data," says P. Stanley Peters Jr. of Stanford University. As a result, linguists have developed an array of competing grammars, all trying to shed light on how human languages work.

One of the earliest grammars to be postulated and studied was based on the observation that in grammatical sentences, certain words tend to appear in clusters. "They are grouped into phrases, and those phrases are broken down into subphrases and so on," says Peters. "Phrase structure is somehow an important part of the language."

These word clusters can be put into categories: noun phrases, verb phrases, adjective phrases and perhaps a few dozen other types. The idea is that all phrases of the same type ought to be completely interchangeable. In such a phrase-structure grammar, "anywhere you have a grammatical sentence with one phrase of a certain type," says Peters, "you can just substitute for it any other phrase of the same type, and the result will be a grammatical sentence."

This concept leads to a grammar that is especially simple mathematically. Known as a "context-free phrase-structure grammar," it consists of a finite list of phrase types and a finite set of rules for piecing together sentences. "You're off to the races," says Peters. "You can generate an infinite set of sentences with a finite set of rules."

But does this simple structure really account for all grammatical features of all human languages? Early on, researchers ran into problems. In practice, for example, it turns out that not dozens but thousands of phrase types have to be identified. In addition, this grammar sometimes requires an unwieldy number of complicated rules.

Into this confusing situation stepped Noam Chomsky of the Massachusetts Institute of Technology. He argued in the 1950s that many of the problems can be overcome by defining a small core of phrase types and adopting new rules allowing "syntactic transformations." These transformations generate other forms of sentences that make up the rest of the language.

Chomsky's theory "overcame some apparent shortcomings of phrase-structure grammar," says Peters, "but it opened the door to too much potential variation in languages." It's like a physicist inventing a theory with seven quarks instead of three, he says. The theory predicts all sorts of funny combinations that don't actually show up.

Chomsky also conjectured that human languages are too complex to fit within either finite-state or context-free grammars, the two simplest types. In the case of a finite-state grammar, this was easy to prove. In such a grammar, sentences are generated by a series of choices made one after the other. To generate a sentence, one picks the first word, then the second, then the third, and so on until the sentence is complete.

Chomsky said that a grammar based on such a system can't handle English sentences like "The bananas in the crate...are yellow." The choice of the word "are" depends not on the words immediately before it but on the word "bananas," which may be separated from "are" by an indefinite amount of material.

Similarly, Chomsky contended that while context-free phrase-structure grammars may succeed in distinguishing the grammatical sentences of some languages, they do so only with unnecessarily complex rules. Although this contention wasn't rigorously proven, most linguistic scholars accepted it and began to focus on transformational grammars.

However, as transformational grammars themselves grew more and more complex, a few linguists began to feel that "some truth had been thrown overboard with the falsehood," says Peters. In particular, several researchers began to look again at phrase-structure grammars to see if they could come up with a theory that was mathematically neater and more tightly

Science Service, Inc. is collaborating with JSTOR to digitize, preserve, and extend access to *Science News*. www.jstor.org

constrained than transformational theory.

In 1982, Pullum and Gerald Gazdar of the University of Sussex in Brighton, England, published a paper reviewing all of the important efforts during the previous 25 years to prove that some languages, including Dutch, English and Mohawk, are not context-free. They concluded that none of the arguments was valid. "This made it once again a live theoretical issue," says Pullum.

These conclusions encouraged Pullum and others to examine context-free phrase-structure grammars more closely. "This theory had survived some pretty powerful empirical testing," says Peters.

It was attractive, says Pullum, because "a context-free language can always be handled fairly efficiently. It's easy to determine whether a sentence is grammatical." This makes such a language amenable to computer processing.

"If you go much beyond context-freeness," says D. Terence Langendoen of the City University of New York Graduate Center, "you get into classes of grammars that are very hard to understand formally. They're quite obscure and unpleasant in a mathematical sense."

Recently, the picture changed. In the August LINGUISTICS AND PHILOSOPHY, two papers show for the first time that there are human languages that cannot possibly be described by a context-free phrase-structure grammar. Stuart M. Shieber of SRI International in Menlo Park, Calif., shows it for a Swiss dialect of German, and Christopher Culy, now with the U.S. Peace Corps in Africa, shows it for Bambara, a language spoken in Mali.

In general, says Pullum, context-free phrase-structure grammars can't define languages in which a "string" of arbitrary length in one part of a sentence must have an exact duplicate elsewhere. Both Shieber and Culy found such a structure or its analog in Swiss German and in Bambara.

Shieber concentrates on a class of subordinate clauses in which a string of verbs may be indefinitely distant from their direct objects. These objects, depending on the verb, may be in the accusative or the dative case. The important point is that when such clauses are constructed, some mental bookkeeping is needed to ensure that there are exactly enough of each type of verb and object. Something in one part of a sentence must correspond exactly to something elsewhere in the sentence.

This example from Swiss German illustrates the principle involved. The clause *... mer d'chind em Hans es huus haend wele laa hälfe aastriiche* can be translated literally as: "... we the children [accusative] Hans [dative] the house [accusative] have wanted let help paint." A smoother English translation turns the clause into: "...we have wanted to let the children help Hans paint the house." Theoretically, in Swiss German there's no limit to the number of objects that must be attached to specific verbs elsewhere in the clause.

"Natural languages can indeed cross the context-free barrier," concludes Shieber.

Culy's demonstration turns on a doubling pattern. In Bambara, certain compound words contain a noun stem repeated twice. It must be the same noun stem each time, and the noun stem may be arbitrarily long. The word *wulu*, for instance, means "dog." The construction *wulu o wulu*, in which the noun is repeated exactly, means "whichever dog."

Bambara also contains the following construction: *wulu + nyini + la = wulunyinina*. This translates as "one who searches for dogs" or as "dog searchers." The construction can be "doubled" to produce *wulunyininanyinina*, which means "one who searches for dog searchers." Combining this construction with the one described in the previous paragraph produces the example that started off this article.

The point is that, theoretically, a speaker can build up words of arbitrary length by way of a doubling scheme. Yet if such words are used in the "o" construction, the nouns on each side of "o" must be the same. "This very free process of redoubling causes the vocabulary of Bambara to be non-context-free," says Culy.

Linguists are still debating whether English contains a similar feature. The sentence "Expert or no expert, I think he's wrong" may fill the bill. The noun "expert," some linguists argue, can be replaced by an arbitrarily long noun phrase. But because the same words must appear on either side of "or no," the situation is similar to that found in Bambara or Swiss German. Not everyone, however, is convinced that this argument is valid.

Nevertheless, to Pullum and others, the evidence now clearly points to the fact that at least two human languages are generally somewhat more complex than the simplest grammars would allow. In the end, Chomsky was right. "It seems now that we have sufficient grounds to believe that this difficult, and for 20 years controversial, issue has been resolved," says Langendoen.

But context-free phrase-structure grammars are hard to give up. Because of their straightforward mathematical structure, context-free languages are easy to analyze. Well-known and highly efficient algorithms do the job nicely. If the recently discovered exceptions prove to be minor, then something may still be salvaged.

"The context-free grammars come very close to describing all structure of all languages," says Pullum. "In fact, the tricky patterns that make for the non-context-freeness, when studied in their own terms, turn out to be fairly easy to describe and likely to be tractable as well.

"Although you can't deal with a language that has redoubling and duplication using a context-free grammar, and therefore there are certain parsing techniques, familiar in computer science, that won't work on a language like that," he says, "nevertheless, you can easily devise another technique that will."

The emphasis on grammar is important, says Langendoen. "To get a system to work effectively with language as we understand it," he says, "we're going to have to have pretty precise models of grammar in place." A computer needs to know more than just the meanings of words.

"I still resist the conclusion that there's anything about natural languages that makes them wild and woolly," insists Pullum. "I think this kind of mathematical attention to the details of language structure is leading to the conclusion that we can deal with the structures of these languages very successfully using computers. I think the age of true natural language processing is on the horizon."

But there are deeper linguistic questions. "The mind is a black box," says Barbara Partee of the University of Massachusetts at Amherst. "We're trying to understand how it works." People may be "hard-wired" for certain kinds of syntax and not for others, she says.

"You say something to me," says Peters. "I say something to you." Within a fraction of a second, he notes, people do what appears to be an unbelievably complicated analysis involving the linguistic structure. "One of the things we would like to understand about language," he says, "is how we're able to make the mental computations that are involved in going from the perception of a blast of noise...to grasping the information that the blast of noise actually represents."

One appeal of context-free phrase-structure grammars was the ease with which they could be processed. "If something like this could really be made to work as a theory of language," says Peters, "then it would also give you a good start on answering the question of how it is all computationally feasible."

That was the hope; the reality is somewhat different. The Culy and Shieber discoveries now show that developing the right grammar is more complicated than it had seemed. □